Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

R.K. Boel, J.H. van Schuppen

Overload control for switches of communication systems -
A two-phase model for call request processing

Department of Operations Research and System Theory Report OS-R8601 January

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

Overload Control for Switches of Communication Systems-

A Two-Phase Model for Call Request Processing

R.K. Boel
Laboratorium voor Theoretische Elektriciteit, Rijksuniversiteit Gent
Grote Steenweg Noord 2, B9710 Gent-Zwijnaarde, Belgium

J.H. van Schuppen
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

In modern telephone networks the switching and connecting operations are performed by computer controlled switches called Stored Program Control (SPC) exchanges. One of the problems with these switches is the severe performance degradation during periods in which the demand for service exceeds the design capacity. The problem of overload control is to decide whether to admit or not to admit a call request such as to maximize the number of successfully completed calls. In this paper a new and rather general model for the switch is proposed in which the delay in processing a call request is modelled by two phases. From this a simplified model is deduced consisting of a series connection of three random delays. The problem of overload control is then formulated as a stochastic control problem. The solution is a bang-bang control, meaning that a customer is either admitted or not admitted to the switch without randomization.

1980 Mathematics Subject Classification: 93E20, 90B22, 60K25.
Key Words & Phrases: overload control, stochastic control, queueing theory, communication systems.
Note: This publication has been submitted elsewhere.

1. Introduction

The purpose of this paper is to show the role of stochastic optimal control methods in designing a control gate in order to prevent congestion in a computer controlled telephone switch. The purpose of the controller is to maximize the long term rate of successful toll-paying connections. The limiting factors are the wastage of capacity in overload situations due to impatient callers terminating their call requests, the lack of memory space and the positive feedback due to retrial call requests from unsatisfied customers. The novelty of this paper is a model for a switch consisting of a flow network and a two phase model for the processing delay. This model is an improvement and extension of earlier models [1, 3, 6, 7, 13, 12, 15, 14, 17, 18]. The overload control problem is formulated as a stochastic control problem that under certain conditions is shown to have a bang-bang solution.

After a calling party has requested a connection by lifting the phone off-hook the telephone switch to which it is connected has to perform a large number of tasks in order to establish the connection with the requested phone. These tasks are: give a dial tone, read in the digits of the destination, check the validity of the destination number, establish a route through the network and through intermediate switches, give a busy or ringing tone etc. In modern electronic switches all these tasks for different call requests running simultaneously are handled by one or more processors. Clearly the different call requests are in contention for the real time of the same processor and for the same processor memory. The more call requests there are the longer the delay in executing the tasks and

Report OS-R8601
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

hence the longer the delay in establishing the connection. This increases the probability that impatient callers will hang up when only part of their tasks have been carried out or that an admitted call request will be shut out for lack of memory space. In each case processor time and memory space have been wasted on tasks that do not lead to a successful connection. It has been recognized a long time ago [23] that this phenomenon can lead to an instability in the sense that an increase in the rate of call requests may lead to a decrease of the rate of successful connections. When this happens one says that there is *congestion.* The *overload control problem* is then to regulate admission to the switch to prevent congestion. A further complication is the positive feedback due to retrials. Unsuccessful call requests have a high probability of reappearing at the switch soon afterwards and this further increases the arrival rate.

The problem of contention for limited resources is similar to the bistability of random access communication channels as in the ALOHA network [10], to thrashing in a multiprocessing computer system with paging [5] and even to runaway speed instability of asynchronous machines [9]. In each case one would like, for maximum profit, to place the normal operating point near the boundary of the instability region. Because of the presence of uncertainty this however decreases the mean time until instability or congestion.

In section 2 a general model for a computer controlled switch is proposed. It consists of a network flow model for call requests and of a model for the delay in processing call requests. The network flow model includes a queue for retrial call requests. In the model for the processing delay one distinguishes two phases. In the first phase a call request generates tasks sequentially. In particular these tasks represent the reception of externally triggered digits. In the second phase tasks are generated iteratively, representing the search for a route. The resulting stochastic dynamic system consisting of a countable state Markov process model admits a dynamic decomposition. From the general model a simplified model is deduced that consists of a series connection of three random delays.

In section 3 the stochastic control problem is posed for both the general and the simplified model of maximizing the rate of successful connections. When the state of the dynamic system is observed then the control is shown to be bang-bang, meaning that it either accepts or rejects customers without randomization. Because the number of call requests in the retrial mode cannot be observed one has to consider a partial observation stochastic control problem. Ways to solve this problem are indicated.

The authors have profited from numerous discussions with dr. F.C. Schoute which they gratefully acknowledge. They also thank the governments of Belgium and The Netherlands which through their cultural exchange agreement have provided financial support for the cooperation of the two authors.

2. Description of the model

In this section a model for overload control is proposed that is sufficiently general to cover many types of computer controlled switches such as a private business exchange (PBX), a stored program control exchange (SPC) with a central processor or a SPC exchange with distributed processors.

The formulation of the model is split up in two parts. In the first part a flow network is defined consisting of an entry gate, a load processor, several other gates and their interconnections. This network leaves undefined the dynamic model for the load of the processor and is thus quite general. In the second part a model is proposed for the load, in particular for the delay in processing a call request. Subsequently the mathematical formulation of the model is summarized as a stochastic dynamic system in the form of a countable state Markov process. Finally a simplified model is deduced from the general model via aggregation.

The terminology of point processes that is used in this paper may be found in [2, 4, 19]. General references on queueing theory are [5, 8, 10, 16].

The set of natural numbers is denoted by $N = \{0,1,2,...\}$, and for $n \in N$ is $N_n = \{0,1,2,...,n\}$.

In the following (Ω,F) is a measurable space. Let U_1 be a class of control processes that is restricted later in the discussion. Assume that for any $\overline{u} \in U_1$ there is a probability measure P_u on (Ω,F). Assume further that there is a σ-algebra family $(F_t, \overline{t} \in T)$ that is assumed to satisfy the "usual

conditions". The fact that this σ-algebra family can be chosen independently of the control process is discussed in [4, 17] and will be taken for granted here. All processes are assumed to be at least adapted to the family $(F_t,\ t \in T)$.

Without mentioning otherwise, all stochastic processes will be assumed to be right continuous with left hand limits. If $X:\Omega \times T \to R$ is such a process then the process $\{X(t-),\ t \in T\}$ is obtained from X by making it left continuous at the jump points. If the process $A:\Omega \times T \to N$ is a counting process and if

$$dA(t) = R(t)dt + dM(t),\ N(0),$$

where $\{M(t),F_t,t \in T\}$ is a martingale and $\{R(t),F_t,t \in T\}$ is adapted, then the process R will be called the *rate process* of A with respect to $\{F_t,t \in T\}$. Note that because R occurs under the integral sign the process $\{R(t-),F_t,t \in T\}$ is also a rate process.

2.1. A flow network for call requests

The network for the flow of call requests is indicated in figure 1.

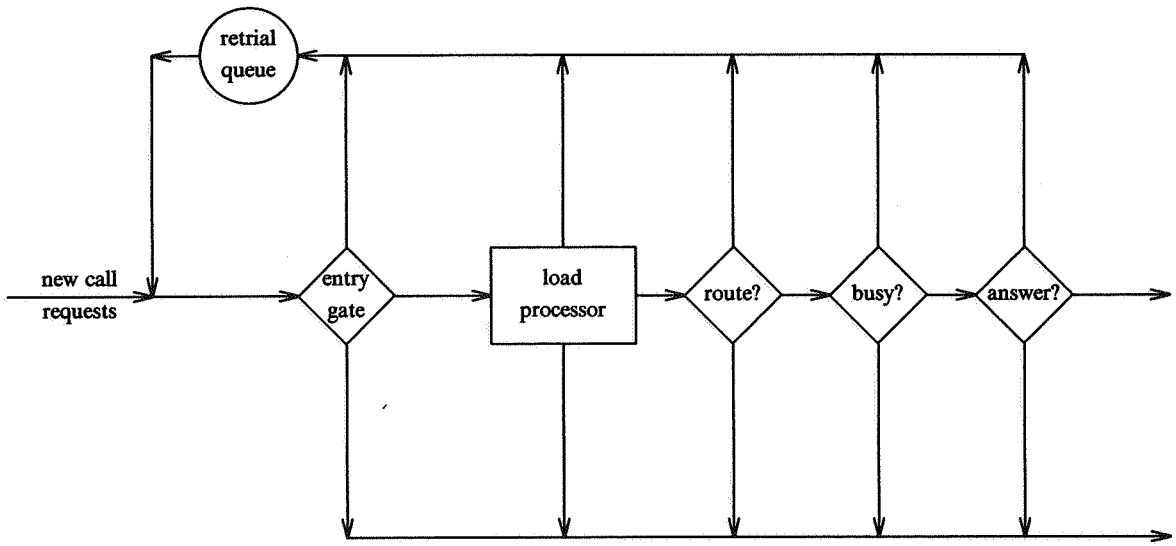

Figure 1. A flow network for a switch.

Arrival of call requests. New call requests from customers are assumed to arrive according to a Poisson process with rate $\lambda_0 \in R_+$.

In addition there are arrivals from the retrial queue. Indeed, some of the callers that request a connection but fail to get it, join the retrial queue and generate another call request after some exponential random time with mean $1/\mu_6$. These random times for different call requests are assumed independent. The arrival rate of retrial call requests is then

$$\mu_6 X_6(t), \tag{2.1}$$

where $X_6:\Omega \times T \to N$ is the number of call requests in the retrial mode.

The entry gate. An incoming call request arrives at the entry gate. This gate either admits the call request to the switch with probability $U(t)$ or rejects the call request with probability $[1-U(t)]$. Note that the control process U takes values in the set $[0,1]$ and must be predictable with respect to $(F_t,\ t \in T)$. In section 3 further restrictions are imposed on the control process.

The call requests that are admitted proceed to the load processor. The call requests that are not admitted either join the retrial queue, with probability r_1 or r_2, depending on whether it is a new or a retrial call request, or leave the switch, with the complementary probabilities. Note that the state of the system does not change if a new call request is rejected and leaves the switch, or if a retrial call request is rejected and again joins the retrial queue.

The load processor. In subsection 2.2 a detailed description is given of the load processor. A call request exits from this submodel by completion of its last task or by a premature departure. In the first case it proceeds to the gate that is labelled "route?" In the latter case it joins the retrial queue with a certain probability or leaves the switch with the complementary probability.

Post-processing. As argued in the introduction a model for overload control has to account for the fact that call requests may after some time reappear at the entry gate because a route is not available, or if a route is available that the requested phone is busy, or if the requested phone is ringing that the call is not answered. A submodel for these events follows.

Assume that a call request leaves the load processor after having completed its last task. The probability that the requested route through the switch and through the network is available is denoted by p_1. If a route is not available, with probability $(1-p_1)$, the call request either joins the retrial queue with probability r_3 or leaves the switch with probability $(1-r_3)$.

If a call request has been given a route it is assumed that the requested phone is not busy with probability p_2. If the requested phone is busy the call request either joins the retrial queue with probability r_4 or leaves the switch with probability $(1-r_4)$.

If a call request has received a ringing tone indicating that the requested phone is not busy, it is assumed that the call request is answered with probability p_3. If the call request is not answered it either joins the retrial queue, with probability r_5, or leaves the switch, with probability $(1-r_5)$.

2.2. A two phase queueing network representing processing delay

A model for the delay in processing call requests is given below. A call request that is admitted to the switch generates tasks. Examples of such tasks are the request for a dial tone, reception of digits of the requested phone number, checks on the correctness of the requested number and the search for a route through the switch and/or through the network. These tasks are executed by the central processor. In a situation of overload or congestion the delay in processing these tasks may be of the same order as the time a customer is willing to wait for service. It is therefore of utmost importance to model both the processing delay and the impatience of customers.

One distinguishes two phases in the processing of call requests. In the first phase a call request must be given a dial tone and the digits of the requested phone must be received and evaluated. During the first phase each call request can at any time have several tasks waiting at the processor. In the second phase, which starts when the destination number has been read in and checked, the search for a route to the requested phone is made. In this phase tasks are generated iteratively; a new tasks being generated only after the preceding task has been executed. Thus in phase two each call request can have at most one task waiting at the processor.

The network of the two phase task processing model is depicted in figure 2.

It is assumed that the buffers of each of the two phases are finite. There are switches in which the buffers for each of the two phases are physically separate, hence it is realistic to assume that the buffer limits may be distinguished. Thus the buffers of the queues 1 and 3 combined are finite with in total X_{13max} places and the buffer of queue 4 is finite with X_{4max} places. A call request is admitted to the switch only if there is space in the buffers of the queues 1 and 3 combined. Hence the

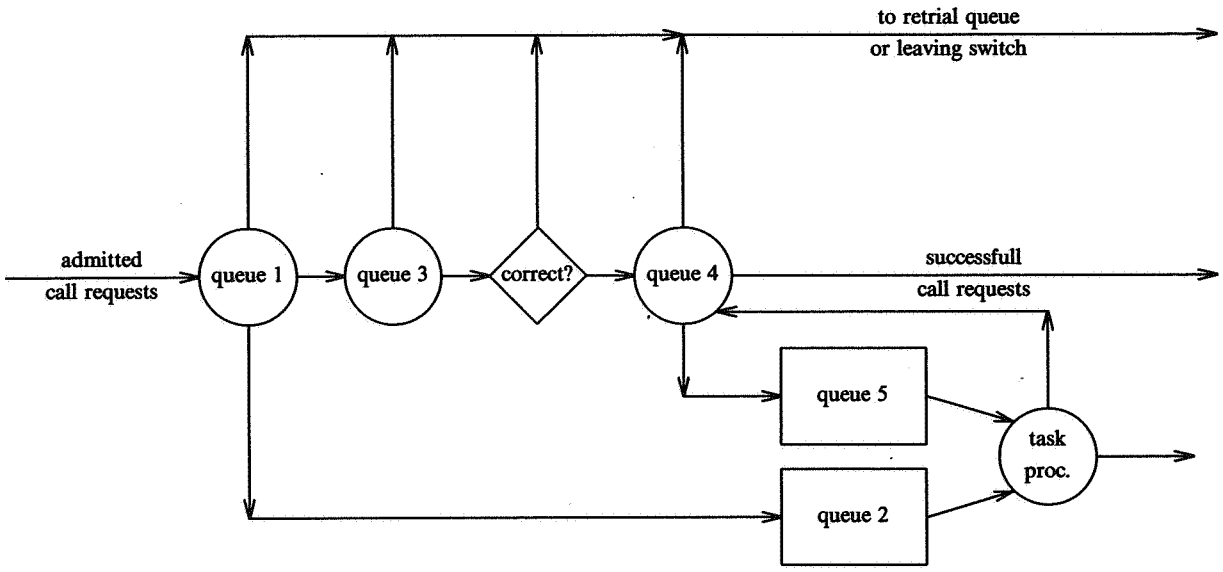

Figure 2. A two-phase model for load processing.

probability of admission, $U(t)$, must be multiplied by

$$I_{(X_1(t)+X_3(t)<X_{13max})} \tag{2.2}$$

If the buffer of phase two is full then the departure rate of queue 2 is put to zero and hence no call request can proceed from phase one to phase two.

Phase one of the processing delay. The subnetwork that models phase one consists of the queues 1, 2 and 3 and a gate, see figure 2. Here queue 1 represents the call requests in phase one that are actively generating tasks. The number of such call requests is denoted by the process $X_1:\Omega\times T\to N$. Queue 3 represents the call requests of phase one that are waiting for the completion of their tasks with $X_3:\Omega\times T\to N$ denoting the number of such requests. Finally queue 2 represents the processing of phase one tasks and $X_2:\Omega\times T\to N$ denotes the number of tasks waiting and being executed. It is assumed that the buffer of queue 2 is finite with X_{2max} places.

The tasks of phase one. Each call request in phase one is assumed to generate a geometrically distributed number of tasks with mean n_1. For different call requests this variable is independent. The time between the generation of two tasks is assumed to be exponentially distributed with mean $1/\lambda_2$ and for different tasks these times are independent random variables. The arrival rate at queue 2 is therefore

$$\lambda_2 X_1(t) I_{(X_2(t)<X_{2max})} \tag{2.3}$$

where the indicator is necessary because tasks are admitted to queue 2 only if the buffer is not full.

The service times of queue 2 are assumed to be exponentially distributed with mean $1/\mu_2$. The service times of different tasks are independent. The service rate is then

$$\mu_2 I_{(X_2(t)>0)} I_{(X_4(t)<X_{4max})}. \tag{2.4}$$

Phase one call requests that generate tasks. Call requests in queue 1 are assumed to actively generate tasks. Because of the assumptions made above, the probability that a generated task is the last task of a particular call request is $1/n_1$. If it is a last task the call request departs from queue 1 and proceeds to queue 3. The departure rate from queue 1 for this transition is

$$\lambda_2 X_1(t)/n_1. \tag{2.5}$$

A call request may also leave from queue 1 because it generates a task while the buffer of queue 2 is full. Such a call request will then join the retrial queue with probability r_6 or leave the switch with the complementary probability. The rates of these two processes are

$$r_6\lambda_2 X_1(t) I_{(X_2(t)=X_{2max})}, \tag{2.6}$$

$$(1-r_6)\lambda_2 X_1(t) I_{(X_2(t)=X_{2max})}. \tag{2.7}$$

Phase one call requests that are waiting for the processing of their tasks. An exact representation of the delay of each individual call request in phase one leads to a complicated model. Therefore an approximation is proposed. A call request departs from queue 3 for one of several reasons.

1. A caller may terminate a call request in the face of a long processing delay or for other reasons. It is assumed that the patience of a customer, or the time he or she is willing to wait, is exponentially distributed with mean $1/\mu_3$. For different customers these random variables are independent. The departure rate for this reason is then

$$\mu_3 X_3(t). \tag{2.8}$$

Note that if the distribution of the time a customer is willing to wait is not exponentially distributed and if there are a relatively large number of call requests present at the queue then the departure rate may be approximated by the formula (2.8). Of this departure process of call requests a fraction r_7 is assumed to join the retrial queue while a fraction $(1-r_7)$ leaves the switch. It is assumed that if a call request departs from queue 3 due to impatience then its waiting phase one tasks will still be executed.

2. A call request may depart from queue 3 because its last task has been completed while it is still present at the switch at the time its last task is completed. The rate at which phase one tasks are processed is given by (2.4). To obtain the defined departure rate, the rate at which phase one tasks are processed must be multiplied by the probability that a completed task is a last task and by the probability that it belongs to a call request that is still present at the switch at the time the last task is completed.

Let $\bar{X}_3:\Omega\times T\to N$ be the stochastic process that denotes the number of last tasks of phase one that are waiting or being served. The probability that a task just completed is a last task is approximated by

$$\frac{\bar{X}_3(t)}{X_2(t)} I_{(\bar{X}_3(t)<X_2(t))} + I_{(\bar{X}_3(t)\geq X_2(t))} = (X_2(t)\wedge\bar{X}_3(t))/X_2(t). \tag{2.9}$$

Next one has to determine the probability that a completed last task belongs to a call request that is still waiting at the time that last task is completed. This probability is approximated by

$$\left(\frac{\mu_2}{\mu_2+\mu_3}\right)^{X_2(t)} = {c_1}^{X_2(t)}, \quad c_1 = \left(\frac{\mu_2}{\mu_2+\mu_3}\right). \tag{2.10}$$

This may be seen as follows. The time S a customer is willing to wait for the processing of phase one tasks from the moment the last task is generated is exponentially distributed with mean $1/\mu_3$. If τ is the time the last task is generated, then the length of queue 2 of phase one tasks at that time is $X_2(\tau)$. The time to complete that particular last task is thus $\sum_{j=1}^{X_2(\tau)} T_j$ where $\{T_j, j\in Z_+\}$ is a col-

lection of independent identically distributed random variables each with an exponential distribution with mean $1/\mu_2$ representing processing times of tasks. The probability that the particular call request is still waiting at the time its last task is completed is then

$$P(\{S \geqslant \sum_{j=1}^{X_2(\tau)} T_j\} \mid F_t^{X_2}), \quad F_t^{X_2} = \sigma(\{X_2(s), s \leqslant t\}). \tag{2.11}$$

If one then approximates $X_2(\tau)$ by $X_2(t)$, because in equilibrium a departing customer on the average leaves behind as many customers as he faced on arrival, then the formula (2.10) is obtained. In this approximation the effect that the buffer of queue 2 or queue 4 may be full is neglected. Note that $\mu_3 \ll \mu_2$.

Finally one approximates

$$X_3(t) \approx \bar{X}_3(t) c_1^{X_2(t)}.$$

The rate of the above formulated departure process is then

$$\mu_2 I_{(X_2(t)>0)} I_{(X_4(t)<X_{4max})} [X_2(t) \wedge (X_3(t)/c_1^{X_2(t)})]\, c_1^{X_2(t)} / X_2(t). \tag{2.12}$$

$$= \mu_2 I_{(X_2(t)>0)} I_{(X_4(t)<X_{4max})} p_5(X(t)), \quad p_5(X(t)) = [X_3(t) \wedge (X_2(t) c_1^{X_2(t)})] / X_2(t).$$

The exit gate of phase one. It is assumed that of all call requests that leave queue 3 because of completion of their last task a fraction $(1-p_4)$ departs because the requested number is incomplete or incorrect. Of this fraction another fraction r_8 joins the retrial queue while the remaining fraction leaves the switch. Finally a fraction p_4 of the call requests leaving queue 3 because of completion of their last task proceeds to queue 4. The fractions of the respective processes are then

$$r_8(1-p_4),\ (1-r_8)(1-p_4),\ p_4.$$

Phase two of the processing delay. The queueing network for phase two consists of queue 4 and queue 5. The call requests in phase two are represented by queue 4, while the number of such call requests is denoted by $X_4: \Omega \times T \to N_{4max}$. Queue 5 represents the tasks of phase two that are waiting for execution by the central processor, while the number of such tasks plus the one being processed is denoted by $X_5: \Omega \times T \to N_{5max}$. It is assumed that the buffer of queue 5 is finite with X_{5max} places.

The tasks of phase two. As stated earlier, in phase two any call request can have at most one task that waits for the processor. It is assumed that a call request in phase two generates a geometrically distributed number of tasks with mean n_2. Furthermore, it is assumed that the time between the completion of the preceding task and the generation of the next task is exponentially distributed with mean $1/\lambda_4$. The rate of the arrival process of phase two tasks at queue 5 is then

$$\lambda_4 (X_4(t) - X_5(t))^+ I_{(X_5(t)<X_{5max})}. \tag{2.13}$$

Here $x^+ = x$ if $x \geqslant 0$, and $= 0$ if $x < 0$. If a task has been completed and the call request is still present then with probability $1/n_2$ that task is considered to be a last task.

The service times of queue 5 are assumed to be independent exponentially distributed random variables with mean $1/\mu_5$. The service rate of queue 5 is then

$$\mu_5 I_{(X_5(t)>0)}. \tag{2.14}$$

The departure process of phase two call requests. A call request departs from queue 4 for one of three different reasons.

1. A task is generated while the buffer of queue 5 is full. In this case the call request is forced to depart from queue 4. It will join the retrial queue with probability r_9 and leave the switch with the complementary probability. The rates of the respective processes are

$$r_9\lambda_4(X_4(t)-X_5(t))^+ I_{(X_5(t)=X_{5max})}, \tag{2.15}$$

$$(1-r_9)\lambda_4(X_4(t)-X_5(t))^+ I_{(X_5(t)=X_{5max})}. \tag{2.16}$$

2. A call request may be terminated because of impatience when the caller is confronted with a long processing delay. The time a customer is willing to wait from the moment its call request enters queue 4 is assumed to be exponentially distributed with mean $1/\mu_4$. For different customers these times are independent. As noted before, this distribution may not be exponential but if there are many call requests waiting their joint rate of departure is approximately proportional to the number of call requests. It is assumed that if a call request departs from queue 4 because of impatience while it has a task that waits in queue 5 then this task will still be executed. Of the call requests that thus depart from queue 4 a fraction r_{10} is assumed to join the retrial queue while the remaining fraction leaves the switch. The rates of these respective processes are

$$r_{10}\mu_4 X_4(t), \tag{2.17}$$

$$(1-r_{10})\mu_4 X_4(t). \tag{2.18}$$

3. Finally a call request may depart from queue 4 because its last task is completed when the call request is still present. The rate at which phase two tasks are processed is given by (2.14). The probability that a call request is still present at the switch at the time one of its tasks is completed may be approximated by

$$\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_5(t)}. \tag{2.19}$$

The explanation of this term is identical to that of (2.10). Note that in the approximation are neglected the contribution of the departure process because the buffer of queue 5 is full and the time necessary to generate a new task. Note that $\mu_4 \ll \mu_5$. If a call request is still present at the moment one of its tasks is completed then it is a last task with approximately probability $1/n_2$.

This departure process from queue 4 then has the rate

$$\frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_5(t)} I_{(X_4(t)>0)} I_{(X_5(t)>0)}. \tag{2.20}$$

2.3. A state space representation

The mathematical models for the flow network and the processing delay will be combined in one state space representation.

The state space is here

$$\underline{X} = \{(k_1,\ldots,k_6) \in N_{13max} \times N_{2max} \times N_{13max} \times N_{4max} \times N_{5max} \times N \mid k_1+k_3 \leq X_{13max}\}.$$

The state process is then

$$X: \Omega \times T \to \underline{X}, \quad X(t) = (X_1(t), X_2(t), X_3(t), X_4(t), X_5(t), X_6(t)).$$

where the processes X_i for $i = 1,2,3,4,5,6$ are as defined before.

Below the dynamics of the Markov process X is summarized. This is done according to an approach given by Walrand and Varaiya [20]. In this formulation one specifies the possible transitions of the Markov process and the rates with which these transitions occur. The transitions are indicated as

$T_i(\textit{old state}) = (\textit{new state})$,

with their respective rate process.

The list of all possible transitions and their respective rate processes is easily deduced from the preceding discussion and given by:

$$T_1(X(t-)) = (X_1(t-)+1, X_2(t-), X_3(t-), X_4(t-), X_5(t-), X_6(t-)), \tag{2.21a}$$

$$R_1(X(t-),(U(t-)) = \lambda_0 U(t-) I_{(X_1(t-)+X_3(t-)<X_{13max})} \tag{2.21b}$$

$$T_2(X(t-)) = (X_1(t-), X_2(t-), X_3(t-), X_4(t-), X_5(t-), X_6(t-)+1), \tag{2.22a}$$

$$R_2(X(t-), U(t-)) = r_1\lambda_0[1-U(t-)I_{(X_1(t-)+X_3(t-)<X_{13max})}], \tag{2.22b}$$

$$T_3(X(t-)) = (X_1(t-)+1, X_2(t-), X_3(t-), X_4(t-), X_5(t-), X_6(t-)-1), \tag{2.23a}$$

$$R_3(X(t-), U(t-)) = \mu_6 X_6(t-) U(t-) I_{(X_1(t-)+X_3(t-)<X_{13max})}, \tag{2.23b}$$

$$T_4(X(t-)) = (X_1(t-), X_2(t-), X_3(t-), X_4(t-), X_5(t-), X_6(t-)-1), \tag{2.24a}$$

$$R_4(X(t-), U(t-)) = (1-r_2)\mu_6 X_6(t-)[1-U(t-)I_{(X_1(t-)+X_3(t-)<X_{13max})}], \tag{2.24b}$$

$$T_5(X(t-)) = (X_1(t-), X_2(t-)+1, X_3(t-), X_4(t-), X_5(t-), X_6(t-)), \tag{2.25a}$$

$$R_5(X(t-)) = \lambda_2 X_1(t-)(1-1/n_1) I_{(X_2(t-)<X_{2max})}, \tag{2.25b}$$

$$T_6(X(t-)) = (X_1(t-), X_2(t-)-1, X_3(t-), X_4(t-), X_5(t-), X_6(t-)), \tag{2.26a}$$

$$R_6(X(t-)) = \mu_2 I_{(X_2(t-)>0)} I_{(X_4(t-)<X_{4max})}[1-p_5(X(t-))], \tag{2.26b}$$

$$T_7(X(t-)) = (X_1(t-)-1, X_2(t-)+1, X_3(t-)+1, X_4(t-), X_5(t-), X_6(t-)), \tag{2.27a}$$

$$R_7(X(t-)) = \lambda_2 X_1(t-) I_{(X_2(t-)<X_{2max})} / n_1, \tag{2.27b}$$

$$T_8(X(t-)) = (X_1(t-)-1, X_2(t-), X_3(t-), X_4(t-), X_5(t-), X_6(t-)+1), \tag{2.28a}$$

$$R_8(X(t-)) = r_6\lambda_2 X_1(t-) I_{(X_2(t-)=X_{2max})}, \tag{2.28b}$$

$$T_9(X(t-)) = (X_1(t-)-1, X_2(t-), X_3(t-), X_4(t-), X_5(t-), X_6(t-)), \tag{2.29a}$$

$$R_9(X(t-)) = (1-r_6)\lambda_2 X_1(t-) I_{(X_2(t-)=X_{2max})}, \tag{2.29b}$$

$$T_{10}(X(t-)) = (X_1(t-), X_2(t-)-1, X_3(t-)-1, X_4(t-), X_5(t-), X_6(t-)+1), \tag{2.30a}$$

$$R_{10}(X(t-)) = r_8(1-p_4)\mu_2 I_{(X_2(t-)>0)} I_{(X_4(t)<X_{4max})} p_5(X(t-)), \tag{2.30b}$$

$$T_{11}(X(t-)) = (X_1(t-), X_2(t-)-1, X_3(t-)-1, X_4(t-), X_5(t-), X_6(t-)), \tag{2.31a}$$

$$R_{11}(X(t-)) = (1-r_8)(1-p_4)\mu_2 I_{(X_2(t-)>0)} I_{(X_4(t-)<X_{4max})} p_5(X(t-)), \tag{2.31b}$$

$$T_{12}(X(t-)) = (X_1(t-), X_2(t-)-1, X_3(t-)-1, X_4(t-)+1, X_5(t-), X_6(t-)), \tag{2.32a}$$

$$R_{12}(X(t-)) = p_4\mu_2 I_{(X_2(t-)>0)} I_{(X_4(t-)<X_{4max})} p_5(X(t-)), \tag{2.32b}$$

$$T_{13}(X(t-)) = (X_1(t-), X_2(t-), X_3(t-), X_4(t-), X_5(t-)+1, X_6(t-)), \tag{2.33a}$$

$$R_{13}(X(t-)) = \lambda_4(X_4(t-)-X_5(t-))^+ I_{(X_5(t-)<X_{5max})}, \tag{2.33b}$$

$$T_{14}(X(t-)) = (X_1(t-), X_2(t-), X_3(t-), X_4(t-), X_5(t-)-1, X_6(t-)), \tag{2.34a}$$

$$R_{14}(x(t-) = \mu_5 I_{(X_5(t-)>0)}[1-\frac{1}{n_2}(\frac{\mu_5}{\mu_4+\mu_5})^{X_5(t-)}], \tag{2.34b}$$

$$T_{15}(X(t-)) = (X_1(t-), X_2(t-), X_3(t-), X_4(t-)-1, X_5(t-), X_6(t-)+1), \tag{2.35a}$$

$$R_{15}(X(t-)) = r_{10}\mu_4 X_4(t-)+r_9\lambda_4(X_4(t-)-X_5(t-))^+ I_{(X_5(t-)=X_{5max})}, \quad (2.35b)$$

$$T_{16}(X(t-)) = (X_1(t-),X_2(t-),X_3(t-),X_4(t-)-1,X_5(t-),X_6(t-)), \quad (2.36a)$$

$$R_{16}(X(t-)) = (1-r_{10})\mu_4 X_4(t-)+(1-r_9)\lambda_4(X_4(t-)-X_5(t-))^+ I_{(X_5(t-)=X_{5max})}, \quad (2.36b)$$

$$T_{17}(X(t-)) = (X_1(t-),X_2(t-),X_3(t-),X_4(t-)-1,X_5(t-)-1,X_6(t-)+1), \quad (2.37a)$$

$$R_{17}(X(t-)) = \frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_5(t-)} I_{(X_4(t-)>0)} I_{(X_5(t-)>0)} \quad (2.37b)$$

$$[(1-p_1)r_3+p_1(1-p_2)r_4+p_1p_2(1-p_3)r_5],$$

$$T_{18}(X(t-)) = (X_1(t-),X_2(t-),X_3(t-),X_4(t-)-1,X_5(t-)-1,X_6(t-)), \quad (2.38a)$$

$$R_{18}(X(t-)) = \frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_5(t-)} I_{(X_4(t-)>0)} I_{(X_5(t-)>0)} \quad (2.38b)$$

$$[(1-p_1)(1-r_3)+p_1(1-p_2)(1-r_4)+p_1p_2(1-p_3)(1-r_5)+p_1p_2p_3],$$

$$T_{19}(X(t-)) = (X_1(t-),X_2(t-),X_3(t-)-1,X_4(t-),X_5(t-),X_6(t-)+1), \quad (2.39a)$$

$$R_{19}(X(t-)) = r_7\mu_3 X_3(t-), \quad (2.39b)$$

$$T_{20}(X(t-)) = (X_1(t-),X_2(t-),X_3(t-)-1,X_4(t-),X_5(t-),X_6(t-)), \quad (2.40a)$$

$$R_{20}(X(t-)) = (1-r_7)\mu_3 X_3(t-). \quad (2.40b)$$

With the above defined transitions and their rate processes one obtains the following representation:

$$dX(t) = \sum_{i=1}^{m}[T_i(X(t-))-X(t-)]\, dN_i(t), \quad X(0), \quad (2.41)$$

$$dN_i(t) = R_i(X(t),U(t))dt + dM_i(t), \quad N_i(0), \quad (2.42)$$

where there are $m=20$ possible transitions and where the $(M_i(t),F_t,t\in T)$ are martingales. The representation (2.41,2.42) is the stochastic dynamic system that represents the dynamic behavior of the switch. Note that the rates R_i are linear functions of the control U.

The above expression may be summarized by

$$dX(t) = [f_1(X(t)) + f_2(X(t))U(t)]dt + dM(t), \; X(0). \quad (2.43)$$

Comments on the general model. The mathematical model for overload control admits a dynamic decomposition. A dynamic decomposition is called a hierarchical decomposition in [5]. The values of the parameters are for a realistic model such that the dynamics of the queues 2 and 5 for tasks are fast with respect to the dynamics of the queues 1, 3 and 4 for call requests which in turn are fast with respect to the dynamics of the retrial queue. A steady state analysis may therefore be performed via near complete decomposibility, see Courtois [5].

The general model contains as special cases those in which the call request processing is done by either only a phase one procedure [3, 13, 17] or only a phase two procedure [18].

The model for overload control contains several parameters that have to be determined. In particular:

1. the values of the parameters p_2,p_3,n_1 and the $\{r_i,i\in\{1,2,...,10\}\}$ are assumed to be constant in time and can be estimated by statistical methods;
2. the values of the parameters p_1 and n_2 are assumed to be slowly varying in time depending on the state of the network, slow compared to the dynamics of the switch;

3. the values of μ_2, μ_5 that model the assignment of the processor to phase one and phase two tasks. Since phase one and phase two tasks are executed by the same processor the service rates μ_2 and μ_5 are obviously related. One assumes that the designer has decided in advance on the state dependent priority given to either phase one or phase two tasks. At least it is required that $\mu_2 + \mu_5 \leqslant \mu$ where μ is the maximum service rate of the processor. The service rate of queue two is put to zero if the buffer of queue 4 is full in order to prevent loss of call requests. This amounts to giving absolute priority to phase two tasks.

2.4. A simplified model

From the general model proposed earlier in this section one may deduce a simplified model using aggregation techniques. One deletes the retrial queue and includes its effect by a slow adjustment of the arrival rate. Next the queues 2 and 5 that represent task processing are deleted from the model. When queue 2 is deleted its effect is averaged out according to the theory of near complete decomposability by averaging the transition rates over the equilibrium distribution of X_2. As for queue 5, one assumes that a new phase two task is generated almost immediately on completion of the preceding phase two task. This amounts to taking $\lambda_4 >> \mu_5$ and hence $X_5 = X_4$. What remains are the queues 1, 3 and 4 for the call requests. To simplify figure 2 one also deletes the exit gate between queue 3 and queue 4. This however does not further simplify the analysis of the model. The resulting network is displayed in figure 3.

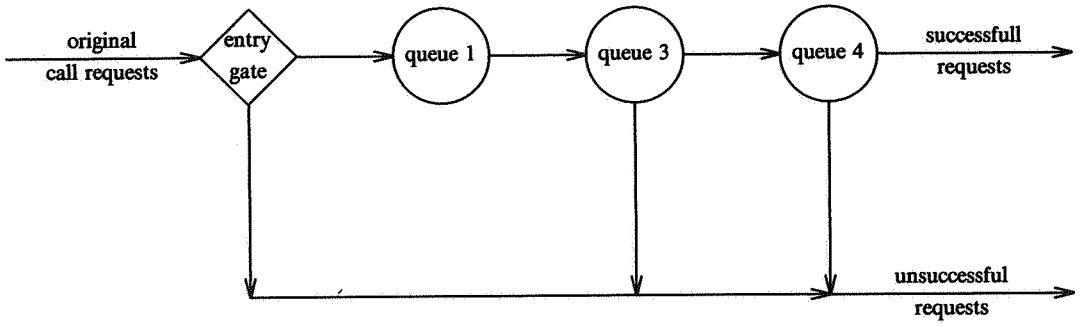

Figure 3. A simplified model for overload control.

For the simplified model one has the following transitions and rates of transitions:

$$T_1(X(t-)) = (X_1(t-)+1, X_3(t-), X_4(t-)), \tag{2.44a}$$

$$R_1(X(t-), U(t-)) = \lambda_0 U(t-) I_{(\lambda_2 X_1(t-) < \mu_2)} I_{(X_1(t-)+X_3(t-) < X_{13max})}, \tag{2.44b}$$

$$T_2(X(t-)) = (X_1(t-)-1, X_3(t-)+1, X_4(t-)), \tag{2.45a}$$

$$R_2(X(t-)) = \lambda_2 X_1(t-)/n_1, \tag{2.45b}$$

$$T_3(X(t-)) = (X_1(t-), X_3(t-)-1, X_4(t-)+1), \tag{2.46a}$$

$$R_3(X(t-)) = \mu_2 I_{(X_4(t-) < X_{4max})} \frac{(1-\rho(t-))}{(1-\rho(t-)^{X_{2max}+1})} \tag{2.46b}$$

$$[\sum_{k=1}^{X_{2max}} \rho(t-)^k [X_3(t-) \wedge (kc_1{}^k)]/k], \quad \rho(t) = \lambda_2 X_1(t)/\mu_2,$$

$$T_4(X(t-)) = (X_1(t-), X_3(t-)-1, X_4(t-)), \tag{2.47a}$$

$$R_4(X(t-)) = \mu_3 X_3(t-), \tag{2.47b}$$

$$T_5(X(t-)) = (X_1(t-), X_3(t-), X_4(t-)-1), \tag{2.48a}$$

$$R_5(X(t-)) = \mu_4 X_4(t-) + \frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_4(t-)} I_{(X_4(t-)>0)}. \tag{2.48b}$$

The stochastic dynamic system that represents the dynamic behavior of the switch is then summarized by:

$$\begin{aligned} \underline{X}_1 &= \{(k_1,k_3,k_4)\in N^3 \mid k_1+k_3 \leqslant X_{13max},\ k_4 \leqslant X_{4max}\}, \\ X(t) &= (X_1(t), X_3(t), X_4(t)), \quad X:\Omega\times T \to \underline{X}_1, \\ dX(t) &= \sum_{i=1}^{5} [T_i(X(t-)) - X(t-)]\, dN_i(t), \quad X(0), \end{aligned} \tag{2.49}$$

$$dN_i(t) = R_i(X(t), U(t))dt + dM_i(t), \quad N_i(0). \tag{2.50}$$

The transitions and the rates of the simplified model follow from the corresponding rates of the general model with the following modifications. By limiting the access of call requests such that $\lambda_2 X_1(t) < \mu_2$ one can assume that queue 2 is in equilibrium with distribution

$$P(\{X_2 = k\}) = \rho(t)^k [1-\rho(t)] \ / [1-\rho(t)^{X_{2max}+1}], \quad k \in N_{X_{2max}}. \tag{2.51}$$

The departure rate of queue 3 due to completed last tasks of still present call requests as given by (2.32b) must then be averaged over the distribution of X_2 from which (2.46b) follows immediately.

While the above simplified model gives only an averaged description of the processing delay of individual tasks, it does include the main causes for congestion except for the retrials. It could be further simplified by replacing, via an approximate aggregation or Norton equivalence [8, 21, 22], the series connection of queue 1 and queue 3 by one single queue with $X_1 + X_3$ call requests of phase one in it. This seems useful for the control problem to be discussed in the next section, since it is quite a reasonable assumption that only the number of call requests present in respectively phase one and phase two will be measured in practice.

3. The stochastic control problem

The dynamic models of section 2 will now allow us to study various methods for preventing congestion by limiting admissions to the switch.

A simple heuristic control could operate as follows. Suppose one imposes a minimum service rate $\bar{\mu}_{1min}$ respectively $\bar{\mu}_{2min}$ for phase one respectively phase two tasks. This leads to a control in which one blocks new admissions when

$$\mu_2 I_{(X_4(t-)<X_{4max})} \frac{(1-\rho(t-))}{(1-\rho(t-)^{X_{2max}+1})} \sum_{k=1}^{X_{2max}} \rho(t-)^k [X_3(t-) \wedge (kc_1^k)] / k \leqslant \bar{\mu}_{1min}.$$

The designer moreover has to assure that

$$\frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_4(t-)} \geqslant \bar{\mu}_{2min}$$

by giving priority to phase two tasks when $X_4(t)$ becomes large.

In the next paragraph of this section one studies the control problem from a different angle. Starting with a reward structure and the average rate of successful toll-paying, customers, one derives the corresponding optimal admission probability. Approximations of the optimal control law will probably achieve a better performance than the heuristic control described above.

3.1. Stochastic optimal control for the general model

The goal of overload control is to maximize the longterm rate of successful call requests. The process of successful call requests is given by

$$dD_S(t) = p_1p_2p_3\frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_5(t)} I_{(X_4(t)>0)}I_{(X_5(t)>0)}dt + dM_s(t), D_S(0). \tag{3.1}$$

The state process is generated by the equations (2.41,2.42) or alternatively by equation (2.43).

The information available to the control is the past of the processes X_1, X_2, X_3, X_4, X_5. Thus all the processes in the switch are observed while the number of retrial call requests is not observed. Let $(G_t, t \in T)$ be the σ-algebra family generated by the observations.

The precise problem formulation proceeds via a measure transformation, see [4, 17]. Thus with

$$\underline{U} = \{U:\Omega\times T\to[0,1] \mid (U(t), G_t, t\in T) \ \textit{predictable}\}$$

for any $U \in \underline{U}$ there exists a probability measure P_u on (Ω, F) such that the state process has the representation (2.41,2.42).

PROBLEM 3.1. Given the stochastic control system (2.41,2.42) with the class of admissable controls $\underline{U}$. Determine a $U^* \in \underline{U}$, if one exists, such that either the *average reward*

$$\lim_{t_1\to\infty} E[D_S(t_1)/t_1] \tag{3.2}$$

$$= \lim_{t_1\to\infty}(1/t_1)E\left[\int_0^{t_1} p_1p_2p_3\frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_5(t)} I_{(X_4(t)>0)}I_{(X_5(t)>0)}dt\right]$$

or the *discounted reward,* for a $c \in (0,\infty)$,

$$E\left[\int_0^{\infty} e^{-ct}p_1p_2p_3\frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_5(t)} I_{(X_4(t)>0)}I_{(X_5(t)>0)}dt\right] \tag{3.3}$$

is maximized.

If the parameters p_1, p_2, p_3, and n_2 are constant over time then they can be deleted from the cost functions. General references on the optimal control of networks of queues are [2, 4, 11, 16, 17].

The state observed case. Assume for a moment that the control is not only predictable with respect to the σ-algebra family $(G_t, t\in T)$ but also with respect to $(F_t, t\in T)$. Note that the stochastic control system (2.43) has the form

$$dX(t) = [f_1(X(t)) + f_2(X(t))U(t)]\,dt + dM(t),\ X(0),$$

and is thus linear in the control while the cost function does not depend explicitly on the control. For the discounted finite-horizon stochastic control problem the result of [3] is applicable and one obtains the result that the optimal control is of bang-bang type. It seems that if there exists at least one control that leads to a stationary distribution then one can show that the stochastic control problems with infinite-horizon discounted cost or average cost also have a bang-bang solution.

The partially observed stochastic control problem. In practice the number of call requests that are in the retrial mode is never observable. So assume that the control U is only predictable with respect to $(G_t, t\in T)$. It is well known in stochastic control theory that a partially observed stochastic control problem must be converted into a state observed stochastic control problem by filtering the unobservable components of the state process. Thus one has to estimate the state component X_6 based on the information in the σ-algebra family $(G_t, t\in T)$. Because the process X_6 takes only countably many values its estimate may be obtained from the family of processes $(\hat{X}_6(t,k), G_t, k\in N, t\in T)$, where

$$X_6(t,k)=I_{(X_6(t)=k)}, \quad \hat{X}_6(t,k)=E[X_6(t,k)\,|\,G_t].$$

An explicit stochastic differential equation for $\{\hat{X}_6(t,k),k\in N\}$ that is driven by the observations can be written [4]. The new state process

$$\{X_1(t),\cdots,X_5(t),\hat{X}_6(t,k),k\in N,t\in T\}$$

then again has a dynamic representation that is linear in the control U, hence it seems that if the optimal control exists it is of bang-bang type. However, this result is not very useful because the control will depend on the infinite family of processes $(\hat{X}_6(t,k),k\in N,t\in T)$. One therefore has to consider an approximation to the retrial queue. This is done below.

Approximation of the retrial queue. Below two approximations for the retrial queue and a selftuning control scheme are proposed.

A diffusion approximation. Approximate the process $X_6(t)$ by the diffusion process $X_7:\Omega\times T\to R_+$. The stochastic differential equation for X_7 may be deduced from the equation for X_6 and has the form

$$dX_7(t) = [f_3(\bar{X}(t))-f_4(\bar{X}(t),X_7(t))U(t)]dt+g(\bar{X}(t),X_7(t))dV(t),X_7(0), \tag{3.4}$$

where

$$\underline{X}_2 = \{(k_1,\ldots,k_5)\in N\times N_{2max}\times N\times N\times N_{5max},\ |\ k_1+k_3\leqslant X_{13max},k_4\leqslant X_{4max}\},$$
$$\bar{X}:\Omega\times T\to\underline{X}_2, \quad \bar{X}(t)=(X_1(t),X_2(t),X_3(t),X_4(t),X_5(t)),$$

where $V:\Omega\times T\to R$ is a standard Brownian motion process and the functions f_3,f_4,g have to be determined. Given (3.4) and an equation for the arrival rate of call requests, $\lambda_0+\mu_6X_7(t)$, one can derive an extended Kalman filter that produces an estimate of the number of retrial mode customers.

A finite state Markov process approximation. The second approximation of the retrial queue is by a finite state Markov process. Abstractly the problem is given a stochastic dynamic system with a countable state Markov process that is driven by and produces a counting process, to determine a stochastic dynamic system with a finite state Markov process that is driven by and produces a counting process such that the input-output behavior of these two stochastic dynamic systems is close in a specified norm.

Stochastic control with restricted observations. In some switches facilities are built in to measure the processes inside the switch. Although in principle one can measure all relevant processes inside the switch, practical considerations suggest to limit the number of measurements.

It is therefore of interest to consider the stochastic control problem for overload control in which the observations consist only of the number of call requests in phase one and phase two, say given the past of the processes X_1,X_3,X_4. This corresponds to the simplified model of subsection 2.4 for which stochastic optimal control is discussed in section 3.3.

3.2. A selftuning adaptive control for the general model

Recall that the dynamics of the retrial queue is slow compared to the dynamics of the queues of the switch. Hence the arrival process is approximately Poisson with a slowly varying rate. The selftuning synthesis procedure of adaptive control suggests using a control law U as if $\lambda_0+\mu_6\bar{X}_6$ is the true arrival rate, where $\bar{X}_6$ is an approximate estimate of the retrial queue length.

In other words, an optimal control is determined that maximizes the rate of successful call requests under the assumption that the arrival rate $\bar{\lambda}$ is known. The optimal control will be of bang-bang type and the switching surface in the state space depends on $\bar{\lambda}$. After some time the value of $\bar{X}_6$ is reestimated and the value $\bar{\lambda}$ and the switching surface are adjusted accordingly. This may lead to a

simple and stable control system.

The same approach may be used if the values of the parameters n_2 and p_1 of the network are slowly changing in time.

3.3. Stochastic control for the simplified model

Consider next the simplified model for overload control of subsection 2.4. The stochastic dynamic system is given by

$$dX(t) = \sum_{i=1}^{5}[T_i(X(t-))-X(t-)]\ dN_i(t),\ X(0), \tag{3.5a}$$

$$dN_i(t) = [R_{i1}(X(t))+R_{i2}(X(t))U(t))]\ dt + dM_i(t),\ N_i(0). \tag{3.5b}$$

The process of successful call requests is given by

$$dD_S(t) = \frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_4(t)} I_{(X_4(t)>0)}\ dt + dM(t),\ D_S(0). \tag{3.6}$$

The class of admissable control policies is

$$\underline{U} = \{U:\Omega\times T\to[0,1] \mid (U(t),F_t^X,t\in T)\ \textit{predictable}\}.$$

PROBLEM 3.2. Determine an optimal control $U^*\in\underline{U}$, if it exists, such that either the average reward

$$\lim_{t_1\to\infty} E\left[\int_0^{t_1}\frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_4(t)} I_{(X_4(t)>0)}dt\right] / t_1, \tag{3.7}$$

or the discounted reward for a $c\in(0,\infty)$

$$E\left[\int_0^{\infty} e^{-ct}\frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{X_4(t)} I_{(X_4(t)>0)}dt\right] \tag{3.8}$$

is maximized.

For the infinite-horizon discounted reward the above problem leads to the Bellman-Hamilton-Jacobi equation

$$-cV(k) + \frac{\mu_5}{n_2}\left(\frac{\mu_5}{\mu_4+\mu_5}\right)^{k} I_{(k>0)} \tag{3.9}$$

$$+ \max_{U\in[0,1]}\left(\sum_{i=1}^{5}[V(T_i(k))-V(k)]\ [R_{i1}(k)+R_{i2}(k)U]\right) = 0,\ k\in\underline{X}_1,$$

where $V:X_1\to R$, $k=(k_1,k_3,k_4)$.

Since the finite set of equations (3.9) has a solution V the optimal control is of bang-bang type and given by

$$U^*(t) = I_{(V(X_1(t-)+1,X_3(t-),X_4(t-)) - V(X_1(t-),X_3(t-),X_4(t-))<0)}. \tag{3.10}$$

The solution to the set of equations (3.9) for V has to be evalutated numerically. Methods for such an approximation remain to be investigated.

The research reported here is part of an ongoing project on overload control of communication systems.

REFERENCES

1. B. BENGTSSON (1982). *On some problems for queues,* Ph.D. thesis" , Linköping University" , Linköping.
2. R.K. BOEL (1985). *Modelling, estimation and prediction for jump processes,* Preprint, Gent.
3. R.K. BOEL and J.H. VAN SCHUPPEN (1985). *Overload control for SPC telephone exchanges - refined models and stochastic control,* Report OS-R8508, Centre for Mathematics and Computer Science, Amsterdam.
4. P. BRÉMAUD (1981). *Point processes and queues - Martingale dynamics*, Springer-Verlag, Berlin.
5. P.J. COURTOIS (1977). *Decomposability: queueing and computer systems applications*, Academic Press, New York.
6. L.J. FORYS (1983). Performance analysis of a new overload strategy. *10th International Teletraffic Congres.*
7. R.L. FRANKS and R.W. RISHEL (1973). Overload model of telephone network operation. *Bell System Techn. J.52*, 1589-1615.
8. E. GELENBE and I. MITRANI (1980). *Analysis and synthesis of computer systems*, Academic Press, New York.
9. C.V. JONES (1967). *The unified theory of electrical machines*, Butterworths, London.
10. L. KLEINROCK (1976). *Queueing systems, volume 2: Computer applications*, John Wiley & Sons, New York.
11. A.A. LAZAR (1983). Optimal flow control of a class of queueing networks in equilibrium. *IEEE Trans. Automatic Control28*, 1001-1007.
12. F.C. SCHOUTE (1983). Adaptive overload control of an SPC exchange. *10th International Teletraffic Congres.*
13. F.C. SCHOUTE (1981). Optimal control and call acceptance in a SPC exchange. *9th International Teletraffic Congres.*
14. F.C. SCHOUTE (1984). *Overload control for the AVS switch,* Report SR-2200-84-4098/19, Dept. SAS, Philips Telecommunicatie Industrie, Hilversum.
15. F.C. SCHOUTE (1983). *The hierarchical queue: A model for definition and estimation for processor loading,* Report SR2200-83-3743, Dept. SAS, Philips Telecommunicatie Industrie, Hilversum.
16. S. STIDHAM JR. (1985). Optimal control of admission to a queueing system. *IEEE Trans. Automatic Control30*, 705-713.
17. J.H. VAN SCHUPPEN (1984). *Overload control for an SPC telephone exchange - An optimal stochastic control approach*, Report OS-R8404, Centre for Mathematics and Computer Science, Amsterdam.
18. B. WALLSTRÖM (1974). On a simple overload control principle for SPC computers. *Proc. 3rd International Seminar on Teletraffic Theory*, 455-466.
19. J. WALRAND and P. VARAIYA (1981). Flows in queueing networks: A martingale approach. *Math. Oper. Res.6*, 387-404.
20. J. WALRAND and P. VARAIYA (1980). Interconnections of Markov chains and quasi-reversible queueing networks. *Stochastic Process. Appl.10*, 209-219.
21. J. WALRAND (1983). A note on Norton's theorem for queueing networks. *J. Appl. Probab.20*, 442-444.
22. J. WALRAND (1983). A probabilistic look at networks of quasi-reversible queues. *IEEE Trans. Information Theory29*, 825-831.
23. J.H. WEBER (1964). A simulation study of routing and control in communication networks. *Bell System Techn. J.57*, 2639-2676.